Cognitive Custody

Access Provenance, Asymmetric Trust, and the Governance of Thought, Files, Identity, Families, and Operational Life-Graphs in AI-Mediated Systems

FIRST EDITION

Elvin D. Almonte, Jr.
Maître d'Œuvre
Maison FORMS

Excerpt from ***The Universal Law of Collapse: A Return to First Principles***

Dominican Republic • Frankfurt — '23-26

© 2023-2026 Maison FORMS LLC. Maison FORMS Holdings LLC. *All rights reserved. This work is protected under U.S. copyright law and international treaties. No part of this document may be reproduced, distributed, or transmitted in any form or by any means without the express written permission of the author. Licensing inquiries via research@maisonforms.com.*

Copyright © 2024-26 Maison FORMS LLC

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or otherwise—without prior written permission of the author.

First Edition

Maison *FORMS™* is a trademark of
Maison *FORMS LLC and Elvin D. Almonte, Jr.*
This work is an original creation of the author and is protected by U.S. and international copyright laws.

Maison FORMS Cultural House
licensing@maisonforms.com

The Laws of The Maison

The Universal Law of Collapse →
Any system extended beyond its foundation will collapse back to that foundation.

The Law of Semantics →
Language is the binding fabric of human cognition: it transforms private perception into shared reality and prevents interpretive drift.

A black box is only as opaque as the information we are permitted to receive.

We must fight not only for the right to identify what is AI-generated, but to protect what is being mapped as cognition-in-progress. Genius once had the luxury of time and the sanctity of the private workshop; today, the tools we depend on strip us of the ability to design ourselves in private.

To the preservation of our most sacred parts: our connections, our ideas, and the quiet orbits of our families.

House Communiqué

from the desk of the
Maître d'Œuvre
Elvin D. Almonte, Jr.

This work was first sharpened by UN Women's article, "When justice fails: Why women can't get protection from AI deepfake abuse." I name UN Women here as an origin point, not as a partner; no endorsement, partnership, or formal relationship is implied. The article gave language to a failure I already recognized: women being harmed faster than institutions can prove, respond, or protect.

The article stayed with me because it described a failure I had already seen. My mother's work and survival have already existed in public record. Domestic violence was not an abstraction in my childhood. I grew up in homeless shelters across four of New York City's five boroughs, and I learned early that protection is not theoretical when the doors around you keep closing.

Maison *FORMS* operates as a house in the old sense: a refuge when other doors close, a shield when truths carry a cost, and a workshop where doctrine becomes tools. I later codified what I had already seen into the House's Universal Law of Collapse: any system extended beyond its foundation collapses back to that foundation. The House was built from that recognition — to stand where foundations are weakest, and to take the hits when telling the truth becomes expensive.

This work begins where women and girls are left exposed, but it does not end there. It widens wherever private life is made visible to systems that owe us proof: families, dependents, creators, inventors, houses, enterprises, and institutions whose lives and work now pass through AI-mediated systems.

AI has entered the workshop, the home, the inbox, the calendar, the archive, and the institution before custody has caught up. The issue is not the machine alone, but the human life now routed through it. A system that can touch a person's images, documents, memories, schedules, relationships, and future locations is not merely a tool. It is a custody environment.

That is why Cognitive Custody was written. Not to accuse every system of malice. Not to reject the usefulness of AI. Not to turn fear into spectacle. It was written because institutions cannot protect what they cannot prove, and people cannot consent meaningfully to access they cannot see.

The first duty is to the vulnerable: to women and girls whose images can be weaponized, to families whose routines can be exposed, to creators whose work can be observed before it becomes public, to inventors whose thoughts may pass through systems before protection is secured, and to institutions asked to govern harms without the primitives required to verify custody, consent, revocation, and response.

This paper is partly a security briefing, partly a governance framework, and partly a demand for a new standard of proof. It asks what must be true if human beings are to create, organize, remember, communicate, and protect one another through systems they do not fully control.

The House offers this work first to those protecting women and girls, and then to every institution willing to treat custody as a duty rather than a slogan. Cognitive Custody is my attempt to name the thing before the world normalizes living beneath it.

The House does not ask for trust without evidence. It asks for custody, provenance, and receipts.

Disclaimer: This paper is not legal advice, cybersecurity advice, or regulatory advice. It is a conceptual and governance framework.

Abstract

This paper argues that AI-mediated systems create an asymmetric custody relationship over cognition-in-progress and operational life-graphs. Large language models, agents, code assistants, cloud connectors, retrieval systems, memory layers, vector databases, open source toolchains, calendars, email accounts, CRMs, repositories, financial systems, and workplace platforms increasingly sit inside the formation layer of human work. They do not merely receive finished artifacts. They observe drafts, prompts, files, schedules, relationships, naming systems, code, invention pathways, business logic, legal reasoning, and operational routines while those things are still becoming.

The paper defines ***cognitive custody*** as third-party access to cognition-in-progress, and ***operational custody*** as third-party access to the connected graph of real-world activity: schedules, communications, files, finances, contacts, tasks, locations, code, obligations, dependents, and decision flows. The central governance demand is ***access provenance***: the right to know who touched the thought, file, identity, family orbit, or operational graph; why access occurred; what was touched; and where those artifacts went. Privacy policies, terms of service, open-source licenses, and compliance promises are insufficient when custody itself remains unverifiable. The proposed remedy is a receipt layer: access receipts, cognitive-chain receipts, operational access receipts, agent execution receipts, connector custody manifests, disclosure-boundary tags, identity-linkage records, and user-facing ledgers.

This is not an anti-AI paper. It is an anti-unverifiable-custody paper.

Executive Summary

The integration of AI into the formation layer of human work—from drafting patents to managing family schedules—has fundamentally altered the boundaries of intellectual and operational privacy. This paper introduces the concepts of ***cognitive custody*** (third-party access to cognition-in-progress) and ***operational custody*** (access to an individual's connected life-graph). It argues that current governance mechanisms, such as privacy policies and terms of service, are structurally inadequate because they rely on unverifiable trust rather than auditable evidence.

To address this asymmetry, this paper proposes a comprehensive, receipt-based governance framework centered on access provenance. This framework demands that AI systems provide verifiable records—such as Cognitive Chain Receipts, Operational Access Receipts, and Non-Use Receipts—detailing who accessed what data, when, why, and how it was used or transformed. By shifting the paradigm from implicit trust to explicit, cryptographic accountability, this framework empowers individuals, creators, and enterprises to maintain sovereignty over their intellectual property and operational lives in an increasingly AI-mediated world.

Introduction: The Workshop is Now Networked

The workshop did not disappear. It became observable. Before the advent of networked AI systems, the traditional intellectual workshop—be it a personal study, a laboratory, or a creative studio—maintained a clear boundary. A physical notebook did not autonomously report its contents, a sketchbook did not generate embeddings for analysis, and a local file directory did not infer the intent behind its organization. A typewriter, a quintessential tool of individual cognition, did not analyze the patterns of its user, nor did a private lab notebook route intermediate reasoning through a complex web of vendors, logs, classifiers, support queues, model providers, or cloud infrastructure. The sanctity of the individual's cognitive process, from nascent thought to finalized artifact, was largely preserved within these physical and localized confines.

The modern AI workflow fundamentally alters this boundary. Individuals now leverage Large Language Models (LLMs) and other AI tools for a vast array of tasks: naming a company, designing an invention, drafting a patent narrative, analyzing a contract, creating a financial model, generating source code, summarizing customer complaints, reviewing medical notes, preparing for meetings, reconciling invoices, or even reflecting on personal crises. In these workflows, the AI system is not merely a passive recipient of finished artifacts; it is positioned upstream, actively observing the questions that precede answers, the discarded pathways that inform decisions, and the emotional or strategic logic that underpins public records. In the AI era, intellectual exposure no longer commences with publication; it begins at the very genesis of cognition-in-progress.

The State of the Field: Synthesis, Not Invention

This paper does not claim that the critical issues of AI privacy, provenance, watermarking, agent tracing, or supply-chain security are entirely unaddressed. Indeed, significant work exists across various domains:

- ***Content Provenance***: Robust systems like C2PA [1] and Content Credentials [2] address the origin and history of digital content, focusing on whether an image or video is AI- generated or has been altered. This supports the paper's claim that content provenance standards are evolving to verify media origin and edits [1, 2].

- ***AI Watermarking***: Technologies such as Google DeepMind's SynthID [3] embed imperceptible watermarks into AI-generated media (images, audio, text) to enable identification. This demonstrates production-ready text watermarking that preserves quality while enabling detection [3].

- ***Agent Tracing***: Research in agent provenance and reasoning traces aims to understand the decision-making processes and actions of autonomous AI agents [4]. This is critical as recent

studies document vulnerabilities in tool-enabled agents, including indirect prompt injection [4, 5].

- ***Software Supply-Chain Security***: Established practices and research focus on securing the code and dependencies that constitute software systems [5].

- ***Privacy Law***: Existing legal frameworks, such as GDPR and CCPA, provide regulations for data protection and individual privacy rights [6].

- ***Access Logs***: Enterprise contexts commonly employ access logs and audit trails to monitor data access within their systems [7].

However, the unique contribution of this paper lies not in the invention of each individual component, but in the ***composition and synthesis of these components into a unified custody framework*** for AI-mediated cognition and operational life-graphs. Existing approaches, while valuable, often operate in silos, failing to address the emergent complexities of an interconnected AI ecosystem. Specifically, this paper bridges critical gaps by focusing on:

- Cognitive Custody: The unique challenges of third-party access to cognition-in-progress.

- ***Operational Custody***: The governance of access to an individual's comprehensive life-Graph.

- ***Life-Graph Exposure***: The holistic risk associated with the aggregation and inference of personal and professional trajectories.

- ***Access Provenance***: A granular, verifiable record of who, what, when, why, and how cognitive and operational data was accessed.

- ***Identity-Linkage Transparency***: Mechanisms to reveal and control how disparate digital identities are correlated by AI systems.

- ***Family/Dependent Exposure***: The specific risks to non-consenting individuals within a user's family orbit.

- ***Receipts for AI-Mediated Access and Non-Use***: A proposed framework for verifiable attestations of data interaction and non-use.

Therefore, this paper's contribution is not the invention of every individual component, but the composition of these components into a custody framework for AI-mediated cognition and operational life-graphs.

Defining Cognitive Custody

Cognitive custody begins where ordinary privacy language becomes too narrow to encompass the complexities introduced by AI-mediated systems. It refers to the condition in which a third-party system possesses access to, exerts influence over, or maintains operational control around an individual's cognition-in-progress. The necessity for this distinct term arises from the inherent limitations of conventional privacy vocabulary. While privacy frameworks primarily address whether a record or piece of data is protected, cognitive custody delves deeper, questioning whether the very formation of thought itself has become visible and auditable by a system whose operations the originator cannot fully verify.

Cognition-in-progress, within this framework, encompasses a broad spectrum of intellectual activity and its digital manifestations. This includes, but is not limited to, prompts, drafts, files, source code, invention trails, research notes, product strategy documents, naming systems, legal reasoning, business logic, emotional reasoning, uploaded documents, connected data, agent logs, memory artifacts, embeddings, summaries, derived metadata, revision pathways, and intermediate outputs. Crucially, this is not merely a collection of static data points. Instead, it represents the dynamic sequence through which intellectual work, from its nascent stages to its refined forms, comes into being. The implications of such pervasive access extend beyond mere data protection, touching upon the fundamental autonomy and intellectual sovereignty of the individual in an AI-augmented world.

The Cognitive Chain

The cognitive chain represents the pre-publication pathway through which a nascent thought or idea evolves into a tangible artifact. In the context of AI-mediated systems, this chain is no longer a solitary, internal process but an intricate sequence that may involve numerous digital touch-points. These can include, but are not limited to, raw thought capture, prompt formulation, model interaction, uploaded files, retrieved contextual information, memory layers, iterative drafting, revision cycles, code generation, agent execution, connector access, cloud routing, system logs, telemetry data, the final artifact itself, and ultimately, publication, filing, or deployment.

While the final output is the visible culmination of this process, its true value often resides within the chain itself. This includes insights into how an idea formed, the alternatives that were considered and discarded, the hidden strategic rationale, the technical pathways that emerged, the boundaries of claims that were explored, the documents that shaped the reasoning, and crucially, which systems had access to this evolving intellectual property before its public disclosure. For domains such as patents, trade secrets, creative authorship, product strategy, and business positioning, the formation path can be as sensitive and valuable as the finished artifact. Thus, the cognitive chain constitutes the pre-publication substrate of intellectual property, demanding a level of scrutiny and governance that traditional frameworks have yet to fully address.

Why Existing IP Frameworks Are Insufficient

Traditional intellectual property (IP) frameworks—comprising copyright, trademark, patent, and trade-secret law—were not conceived in an era where AI systems could observe and influence cognition before the artifact itself materialized. Consequently, these frameworks exhibit significant limitations when confronted with the complexities of AI-mediated intellectual work.

Copyright law, while robust in protecting expression, offers less protection for raw ideas, methodologies, systems, workflows, or reasoning pathways. Its primary focus is on the tangible manifestation of creativity, not the nascent stages of thought. ***Trademark law*** safeguards source identity and prevents consumer confusion, but it does not extend to the underlying conceptual architecture or the developmental process of an idea. ***Patent law*** protects claimed inventions post-filing or disclosure; however, the pre-filing pathway, where critical ideation and development occur, may already have been exposed through AI-mediated workflows, potentially compromising novelty or inventiveness before formal protection can be sought. ***Trade-secret doctrine*** hinges on the maintenance of reasonable secrecy measures, a concept significantly complicated by AI systems where users often lack verifiable insight into who or what accessed their information during the cognitive chain.

The challenge extends beyond mere infringement. A competitor, for instance, does not necessarily need to copy exact text to absorb the structural essence of an idea. They can rename mechanisms, reframe use cases, alter interfaces, shift claim boundaries, or repackage functional equivalents under different optics. Existing IP law, in its current form, is better equipped to protect finished artifacts than the intricate cognitive chain that produced them, leaving a critical gap in the governance of intellectual property in the AI age.

The Asymmetry Problem

The core challenge in AI-mediated cognition lies in a fundamental asymmetry of visibility between users and platforms. Platforms possess comprehensive knowledge regarding access rules, data retention behaviors, internal review workflows, logging practices, model improvement boundaries, safety escalation policies, legal disclosure processes, third-party routing mechanisms, connector behaviors, administrative access pathways, telemetry practices, and exceptions to data deletion protocols. This granular understanding of data flow and access within their systems grants platforms a significant informational advantage.

Conversely, users frequently lack verifiable insight into critical aspects of their data's journey. They often remain unaware of who or what accessed their files, whether such access was human or automated, the precise rationale behind the access, the exact scope of information touched, whether embeddings were created, if summaries were stored, whether logs were retained, or if their data was transferred to other systems. This informational disparity is not incidental; it is structural. While users contribute their cognition—their thoughts, ideas, and operational data—

the platforms retain control over the visibility and audit-ability of how that cognition is handled. Even when a platform operates with the best intentions, the user is unable to independently observe or verify the actions taken with their data. This necessitates a shift from reliance on invisible trust to a framework built on verifiable evidence, particularly as AI-mediated cognition becomes increasingly sensitive and integral to human and institutional life.

Legal Permission vs. Ethical Custody

The distinction between legal permission and ethical custody is paramount in the evolving landscape of AI-mediated systems. Legal permission delineates the boundaries of what a platform may do, as stipulated by terms of service, privacy policies, enterprise agreements, open-source licenses, connector authorization screens, safety exceptions, and various legal compliance obligations. These instruments grant platforms the right to process, access, retain, or disclose data under specific conditions. However, the existence of legal permission does not inherently resolve the question of what a platform should disclose to an individual whose cognition, labor, relationships, or invention pathways it holds.

Ethical custody, in contrast, demands a higher standard of transparency and accountability. It necessitates visibility into data handling practices, proportionality in data access, minimization of data retention, and the implementation of meaningful consent mechanisms. Key components of ethical custody include user-facing logs that detail access events, clear revocability of permissions, disclosure of downstream data usage, explicit clarity on data retention policies, verifiable proof of data interactions, and robust mechanisms for contesting linkages or reviews. While legal permission establishes the permissible actions of a platform, ethical custody defines the imperative for disclosure when a platform assumes stewardship over another person's cognitive output, intellectual labor, or creative processes. This ethical imperative transcends mere legal compliance, advocating for a framework where trust is not assumed but demonstrably earned through transparent custodial practices.

LLMs as Cognitive Infrastructure

Large Language Models (LLMs) have transcended their initial role as mere chatbots to become foundational cognitive infrastructure across a multitude of domains. Their application now spans writing, coding, strategic planning, legal drafting, research, therapeutic reflection, business planning, patent ideation, product architecture, data analysis, creative development, education, enterprise workflows, negotiation preparation, financial reasoning, and decision support. This pervasive integration positions LLMs not just as tools for generating finished outputs, but as active participants in the very formation layer of human cognition.

Crucially, these systems observe the entire sequence of intellectual development, not merely the final artifact. This includes the initial question, moments of hesitation, iterative refinements, architectural considerations, exploration of alternatives, subsequent revisions, and the eventual

articulation of a final claim. Such a comprehensive view can inadvertently reveal a user's underlying assumptions, areas of weakness, strategic ambitions, inherent risks, and private strategies. While a polished white paper typically conceals the dead ends and exploratory detours, a model-mediated drafting session may retain a complete record of them. Consequently, the LLM is not simply processing content; it is deeply embedded at the formation layer of cognition, making its role in intellectual processes far more profound and impactful than previously understood. For instance, a lawyer using an LLM to draft privileged legal strategy inadvertently exposes the iterative development of their arguments, including discarded lines of reasoning and sensitive case details, to the AI system. Similarly, a founder uploading pre-filing patent notes to an LLM for summarization or refinement risks exposing novel intellectual property before formal protection can be secured. These examples underscore how AI systems, by observing cognition-in-progress, become privy to the most sensitive and strategic aspects of human intellectual endeavor.

Access Without User-Visible Disclosure

AI platforms are increasingly entrusted with processing highly sensitive personal, professional, creative, and enterprise-level context. Despite this critical role, users typically lack access to a comprehensive and transparent access ledger. It remains challenging for individuals to ascertain whether a specific chat, file, memory artifact, or connector result has been accessed by an automated classifier, a human reviewer, a support agent, an enterprise administrator, a legal team, an abuse investigation workflow, a model-improvement process, a third-party tool, or through an external disclosure pathway.

While published privacy controls and retention documentation are important, policy disclosure alone does not equate to user-specific access provenance. The fundamental user-facing question remains: who accessed the chat or file, when did this access occur, what was the underlying reason, what was the scope of the access, by what authority was it granted, was the access human or automated, was the data retained, was it transformed, was it disclosed, and was it linked to other identities or systems? Without clear, verifiable answers to these questions, the promise of privacy and control remains largely theoretical, underscoring the urgent need for robust access provenance mechanisms.

Human Access Inside Machine Systems

The discourse surrounding AI systems often portrays them as purely automated entities, operating autonomously without human intervention. However, this perspective overlooks a critical reality: automated systems are invariably embedded within institutional structures and are subject to human oversight and interaction. These institutions typically comprise a diverse array of human actors, including support staff, trust and safety reviewers, contractors, enterprise administrators, legal teams, cloud vendors, incident response teams, evaluation teams, and

platform engineers. Consequently, a system marketed as fully automated may still expose sensitive cognition-in-progress to human review or operational access pathways.

It is important to acknowledge that human access is not inherently illegitimate. Platforms may require human intervention for a variety of valid reasons, such as ensuring safety, preventing fraud, addressing abuse, facilitating account recovery, providing support, or complying with legal mandates. Nevertheless, such human access should not remain invisible to the user. When a file or cognitive artifact crosses the threshold into human review, it fundamentally changes its custody status; it is no longer solely processed by machines but becomes subject to human interpretation and action. For example, if an enterprise connects Slack, Drive, GitHub, CRM, and email to an AI system, and a support agent or internal auditor accesses a user's data for a specific inquiry, this human access, even if authorized, fundamentally alters the custody boundary. Transparency regarding these human touch-points is essential for establishing verifiable custody and maintaining user trust in AI-mediated systems.

Safety Review and the Governance Dilemma

Safety review presents one of the most challenging governance dilemmas within AI-mediated systems, primarily because certain forms of access may be legitimately justified. Scenarios such as imminent harm, child safety concerns, fraud prevention, malware detection, platform security, legal compliance, emergency escalation, account compromise, and misuse prevention often necessitate that systems detect, escalate, or investigate user activity. A robust custody framework must acknowledge and accommodate these critical exceptions.

However, the imperative for safety does not negate the fundamental principles of custody. While safety concerns may alter the justification for access, they do not eliminate the need for transparency and accountability. This includes requiring clear reason codes for access, strict scope limitations, internal audit trails, a clear separation between safety review processes and product learning mechanisms, defined retention boundaries, independent oversight, and, where legally permissible, user notification. Safety exceptions should not serve as a pretext for invisible, general-purpose access. While safety may justify access, it does not absolve platforms of their custodial responsibilities; rather, it reframes them within a context of heightened scrutiny and explicit governance.

Derived Data and Vector Shadows

In AI-mediated systems, the data generated from user interactions extends far beyond the explicit inputs provided. This ***derived data***—including embeddings, summaries, metadata, and other transformed representations—creates what can be termed ***vector shadows***. These shadows are not merely copies of original content but are computationally generated artifacts that capture semantic meaning, relationships, and patterns from the user's cognition-in-progress and

operational life-graphs. While often presented as anonymized or aggregated, these derived representations can, in practice, be highly sensitive and re-identifiable.

The creation of vector shadows introduces new custody challenges. Users typically lack visibility into the generation, storage, and utilization of these derived data forms. They often do not know: (1) what specific algorithms or models were used to create the embeddings or summaries; (2) where these vector shadows are stored and for how long; (3) who has access to them, both human and automated; (4) whether they are used for model training, product improvement, or other undisclosed purposes; and (5) whether they can be linked back to the original user or other identities. The existence of these persistent, yet often invisible, derived data artifacts means that even if original inputs are deleted, a significant and potentially revealing trace of the user's cognitive and operational activity may persist, creating a persistent custody gap that current privacy frameworks struggle to address. For instance, a vector database might retain semantic representations (embeddings) of a user's sensitive documents even after the original files have been deleted from the system. These vector shadows, while not the original data, can still be vulnerable to inversion attacks that reconstruct the original content, thereby exposing information the user believed was erased.

The Non-Use Problem

The "non-use problem" highlights a critical gap in current AI governance: the difficulty in verifying claims of non-use regarding user data. Platforms frequently assert that user data, particularly cognition-in-progress, will not be used for model training, product improvement, or other secondary purposes. However, without robust, verifiable mechanisms, these claims remain largely trust-based rather than evidence-based. The challenge lies in the inherent opacity of AI systems and the complex data pipelines that underpin them.

Users lack the ability to independently audit whether their prompts, drafts, or operational data, even if anonymized or aggregated, are indeed excluded from training sets or other internal processes. This creates a significant custody blind spot. The non-use problem is exacerbated by the economic incentives driving AI development, where vast quantities of data are often seen as essential for model refinement and competitive advantage. A truly transparent custody framework must move beyond mere policy statements to provide users with verifiable receipts or attestations that confirm when their data has not been used for specified purposes, thereby converting a trust claim into an inspectable custody fact. For example, a user might be assured that their prompts are not used for model training, but without a verifiable non-use receipt, they cannot independently confirm this. This is particularly relevant for sensitive data where the expectation of non-use is critical, such as medical records or legal documents.

The Connector Problem: Entangled Data Management

The proliferation of connectors in AI-mediated systems introduces a significant challenge to cognitive custody, creating a complex web of entangled data management. Connectors, which link AI platforms to external services such as email, calendars, CRM systems, cloud storage, and enterprise applications, are designed to enhance functionality and streamline workflows. However, this interoperability often comes at the cost of clear data boundaries and verifiable custody.

When an AI system is granted access to external data sources via connectors, the user's cognition-in-progress and operational life-graphs become distributed across multiple, often disparate, platforms. This entanglement makes it exceedingly difficult for users to track: (1) which specific data points are accessed by which connector; (2) the scope and duration of that access; (3) how data is transformed or processed by the connector before being ingested by the AI system; (4) whether the connector itself retains copies of the data; (5) the security posture and privacy policies of each third-party service involved; and (6) the downstream implications of data sharing across these interconnected systems. The connector problem transforms custody from a single-platform concern into a multi-party, multi-system challenge, where the weakest link in the chain can compromise the entire cognitive and operational graph. Effective governance in this context requires not just platform-level custody receipts, but also comprehensive connector custody manifests that provide granular transparency into data flow and access across the entire entangled ecosystem. For example, an enterprise connecting Slack, Drive, GitHub, CRM, and email to an AI system creates a complex data entanglement. Each connector, while seemingly innocuous on its own, contributes to a holistic operational life-graph that can be accessed and analyzed by the AI. This interconnectedness means that a malicious prompt injection hidden in a connected document or calendar invite could potentially leverage access across all these integrated systems, highlighting the amplified risk of connector-based data exposure.

Operational Custody and the Life-Graph Problem

Beyond the realm of cognition-in-progress, AI-mediated systems increasingly exert ***operational custody*** over an individual's entire operational life-graph. This concept refers to third-party access to the interconnected web of real-world activities that constitute an individual's daily existence. This includes, but is not limited to, schedules, communications, files, financial transactions, contact networks, tasks, geographical locations, code repositories, obligations, and the intricate decision flows that govern personal and professional life.

The ***life-graph problem*** arises because AI systems, through their pervasive integration into digital tools and services, can reconstruct a comprehensive and dynamic map of an individual's operational reality. This goes beyond mere data collection; it involves the aggregation and analysis of disparate data points to infer patterns, predict behaviors, and understand the

underlying structure of a person's life. For instance, an AI system with access to a user's calendar, email, and financial records can not only track past activities but also anticipate future engagements, identify critical relationships, and even deduce financial vulnerabilities. The custody challenge here is multifaceted: (1) the sheer volume and diversity of data points involved; (2) the inferential capabilities of AI that can derive sensitive insights from seemingly innocuous data; (3) the lack of user control or visibility over the reconstruction process; and (4) the potential for this life-graph to be exposed, misused, or targeted. Effective operational custody demands verifiable provenance over every interaction with this life-graph, ensuring that users retain sovereignty over their digital existence. For instance, a breached AI workspace might expose not just individual files, but the entire operational life-graph of a user, revealing their aliases, writing style, company plans, calendar patterns, collaborators, pending deals, file structure, codebase, financial pressure, family routines, and even their future whereabouts. This transformation from a data breach to a life-graph breach represents a critical escalation in the potential harm and necessitates a re-evaluation of security protocols and custodial responsibilities.

Schedule, Relationship, Financial, and Institutional Exposure

The comprehensive nature of operational custody extends to exposing highly sensitive aspects of an individual's life across multiple vectors: schedules, relationships, finances, and institutional affiliations. Each of these areas, when mediated by AI systems, presents distinct and significant custody risks.

Schedule Exposure: AI systems with access to calendars, travel itineraries, and communication logs can reconstruct an individual's past, present, and future movements and commitments. This granular insight into daily routines, appointments, and travel plans creates a detailed temporal map, revealing patterns of presence and absence that can be exploited for various purposes, from targeted advertising to physical surveillance. The ability to predict an individual's location or availability based on their digital schedule constitutes a profound erosion of personal autonomy and security.

Relationship Exposure: The interconnectedness of digital communication platforms, social networks, and collaborative tools allows AI systems to map an individual's relational graph. This includes identifying close contacts, professional networks, familial ties, and even casual acquaintances. By analyzing communication frequency, content, and shared activities, AI can infer the strength, nature, and importance of these relationships. Such exposure can lead to social engineering attacks, reputational damage, or the manipulation of social dynamics, as adversaries gain insight into an individual's most valuable connections.

Financial Exposure: AI integration into banking, investment, and e-commerce platforms grants systems access to an individual's complete financial life-graph. This encompasses spending

habits, income sources, debt obligations, investment portfolios, and transactional histories. The aggregation of this data allows for sophisticated financial profiling, revealing economic vulnerabilities, purchasing power, and potential leverage points. The custody risk here is not merely about data breaches but about the potential for AI to infer and expose an individual's entire economic standing and future financial trajectory.

Institutional Exposure: For individuals operating within organizational structures, AI-mediated workflows can expose their institutional footprint. This includes their roles, responsibilities, project involvements, access privileges, and internal communications within companies, academic institutions, or governmental bodies. AI systems can map an individual's influence, access to sensitive information, and strategic importance within an organization. This institutional exposure carries significant risks for corporate espionage, insider threats, or the compromise of organizational integrity through the targeting of key personnel. In each of these dimensions, the lack of verifiable custody over AI-mediated data transforms routine digital interactions into potential vectors for profound personal and professional vulnerability.

Future-State Exposure

One of the most profound and often overlooked dimensions of cognitive and operational custody is ***future-state exposure***. This refers to the capacity of AI-mediated systems to infer, predict, or even influence an individual's future actions, intentions, and trajectories based on their cognition-in-progress and operational life-graphs. Unlike traditional data collection, which primarily captures past and present states, AI's analytical capabilities extend to probabilistic forecasting, transforming current data into predictive insights about what an individual ***will*** do, ***will*** think, or ***will*** become.

This exposure manifests in several critical ways: (1) ***Predictive Behavioral Modeling***: AI systems can analyze patterns in prompts, drafts, schedules, and communications to predict future decisions, career moves, purchasing behaviors, or even personal relationships. This creates a digital shadow of an individual's potential future, which can be leveraged by platforms, advertisers, or even malicious actors. (2) ***Intent Inference***: The iterative nature of AI interaction, where users refine prompts and explore ideas, allows systems to infer underlying intentions and goals before they are fully articulated or acted upon. This preemptive understanding of intent represents a significant erosion of personal agency and strategic advantage. (3) ***Trajectory Engineering***: In more advanced scenarios, AI systems, particularly those integrated with agents and operational tools, can actively shape an individual's future trajectory. By presenting curated information, nudging decisions, or automating actions based on inferred future states, these systems can subtly guide users towards predetermined outcomes, often without explicit user awareness or consent. The custody challenge of future-state exposure lies in the fact that it operates on the un-manifested, making it exceptionally difficult for users to audit, control, or even perceive the ways in which their potential futures are being observed, predicted, and potentially influenced by AI. This necessitates a framework that extends provenance not just to

past and present data, but to the probabilistic future states derived from it. For instance, an AI system analyzing a user's calendar, communications, and financial data might infer an upcoming career change or a significant personal investment. This inferred future state, if compromised, could be used for targeted manipulation or exploitation, even before the user has fully committed to the action. The custody challenge here is that the AI is not merely reflecting existing data but generating probabilistic insights about un-manifested intentions, making it a powerful tool for predicting and potentially influencing an individual's future.

Family, Dependent, and Household Safety Exposure

AI-mediated systems, through their deep integration into personal and household management, introduce a critical and often overlooked dimension of custody risk: ***family, dependent, and household safety exposure***. When AI tools access or manage information related to family members, children, or household routines, they create a nexus of sensitive data that can have profound implications for the safety and privacy of individuals who have not explicitly consented to such data collection or processing.

This exposure can manifest in several ways: (1) ***Schedule and Location Tracking***: AI systems managing family calendars, school schedules, pediatric appointments, or shared location data can inadvertently create a detailed operational map of a household's movements and routines. This information, if compromised, could be exploited by malicious actors for targeting, surveillance, or even physical harm. (2) ***Dependent Vulnerability***: Information pertaining to children's activities, health records, educational progress, or online interactions, when processed by AI, can expose minors to risks without their informed consent. The unique vulnerabilities of dependents necessitate heightened custodial standards to prevent their data from being used for profiling, exploitation, or unauthorized access. (3) ***Household Operational Patterns***: AI-powered smart home devices, shared digital assistants, and integrated household management systems collect data on daily routines, energy consumption, communication patterns, and even internal conversations. This aggregated data can reveal intimate details about household occupancy, habits, and vulnerabilities, creating a comprehensive digital footprint of the domestic sphere. The challenge of family, dependent, and household safety exposure underscores the need for a custody framework that extends beyond the individual user to protect all members of a connected household, ensuring that privacy and safety are paramount in the design and deployment of AI-mediated services. For example, a parent connecting their Gmail or Calendar to an AI assistant might inadvertently make school pickup schedules, pediatric appointments, or family vacation plans visible to the AI system. If this information is compromised or misused, it could expose children and other dependents to significant safety risks, highlighting how seemingly innocuous data points can compose into a map of future vulnerability for an entire family.

The Composition Problem: Why Isolated Permissions Fail

The prevailing model of digital permissions, characterized by isolated, granular grants for individual applications or data types, proves fundamentally inadequate in the context of AI-mediated cognition and operational life-graphs. This ***composition problem*** arises because AI systems do not operate on discrete, siloed pieces of information; rather, their power derives from the ability to synthesize, correlate, and infer insights from a vast and heterogeneous array of data sources. Consequently, permissions granted in isolation, while seemingly benign, can compose into a far more comprehensive and revealing picture than any single grant suggests.

Consider a scenario where a user grants an AI tool permission to access their calendar (for scheduling), their email (for communication context), and a cloud storage service (for document drafting). Individually, each permission appears limited and justifiable. However, when these permissions are composed, the AI system gains the capacity to: (1) correlate meeting invitations with email discussions and document revisions, thereby reconstructing the entire cognitive chain of a project; (2) infer strategic priorities by analyzing the frequency and content of communications related to specific calendar events; and (3) predict future actions or intentions based on the combined patterns across all three data streams. The user, having granted each permission separately, may remain unaware of the emergent, holistic understanding the AI system has developed.

This compositional capability renders traditional, isolated permission models insufficient for ensuring genuine cognitive custody. The sum of individually granted permissions often far exceeds the user's intended scope of access, creating a systemic vulnerability where the AI can construct a detailed life-graph without explicit, holistic consent. Addressing this requires a shift from atomized permission grants to a more integrated approach that accounts for the combinatorial power of AI, demanding transparency and control over the composition of data access, not just its individual components.

Task-Scoped Consent and Operational Access Receipt

To address the limitations of traditional permission models and the composition problem, this paper proposes the implementation of ***task-scoped consent*** coupled with ***operational access receipts***. Task-scoped consent represents a paradigm shift from broad, blanket permissions to granular, context-specific authorizations. Instead of granting an AI system general access to an entire data repository, consent would be explicitly tied to a particular task or workflow, defining precisely what data can be accessed, for what purpose, and for what duration.

For example, if an AI agent is tasked with summarizing a specific document, task-scoped consent would limit its access solely to that document and the necessary processing resources, preventing

it from browsing unrelated files or accessing other personal data. This approach ensures that the scope of AI access is always commensurate with the immediate operational need, thereby minimizing unnecessary exposure of cognition-in-progress and operational life-graphs.

Complementing task-scoped consent are ***operational access receipts***. These receipts serve as verifiable, immutable records of every instance of AI access to sensitive data or systems. An operational access receipt would document: (1) the specific AI system or agent that initiated the access; (2) the precise timestamp of the access event; (3) the exact data or system components that were accessed; (4) the specific task for which the access was granted; (5) the authority under which the access occurred (e.g., user-initiated, automated workflow, safety review); (6) whether the access was human or automated; (7) the retention policy applied to the accessed data; and (8) any transformations or disclosures of the data. These receipts would be user-facing, providing individuals with a transparent and auditable ledger of all AI interactions with their digital lives. By combining task-scoped consent with operational access receipts, users can gain unprecedented control and visibility over their cognitive and operational custody, transforming implicit trust into explicit, verifiable accountability.

Custody Tiers and Sensitivity Classes

To effectively manage the diverse range of cognitive and operational data entrusted to AI-mediated systems, a nuanced framework of ***custody tiers and sensitivity classes*** is essential. A one-size-fits-all approach to data governance is insufficient, as the implications of exposure vary dramatically depending on the nature and context of the information. This framework proposes categorizing data based on its inherent sensitivity and the potential impact of its compromise, thereby enabling proportionate security measures and disclosure protocols.

Custody Tiers would define the level of protection and audit-ability required for different categories of data. These tiers could range from a baseline level for general, non-sensitive information to highly restricted tiers for privileged or critical data. For instance:

- ***Tier 1 (General)***: Publicly available information, non-sensitive drafts, or data with minimal personal or strategic value. Requires basic access provenance and retention policies.

- ***Tier 2 (Confidential)***: Internal documents, early-stage ideas, or personal communications not intended for public disclosure. Demands enhanced access logs, clear use limitations, and user-facing disclosure of automated processing.

- ***Tier 3 (Sensitive/Proprietary)***: Trade secrets, pre-filing patent ideas, unreleased product roadmaps, financial data, or personal health information. Requires stringent access controls, explicit human-review disclosure, no default training use, and robust audit trails.

- ***Tier 4 (Privileged/Critical)***: Legally privileged communications, national security information, or data whose compromise could lead to catastrophic personal or institutional harm. Necessitates end-to-end encryption, local-first processing options, independent oversight, and immediate user notification of any access event.

Sensitivity Classes would further refine these tiers by categorizing the type of information, such as financial, medical, legal, familial, or strategic. This classification allows for the application of domain-specific regulations and ethical considerations. For example, medical data, even if classified as Tier 3, would be subject to additional HIPAA-like compliance requirements, while legal data would invoke attorney-client privilege considerations.

By implementing distinct custody tiers and sensitivity classes, AI platforms can move beyond generic privacy policies to provide a granular, context-aware approach to data governance. This framework ensures that the level of protection afforded to an individual's cognition-in progress and operational life-graphs is commensurate with its inherent value and the potential risks associated with its exposure, thereby fostering a more secure and trustworthy AI ecosystem.

Privileged Contexts and Professional Custody

The integration of AI into professional workflows introduces acute challenges regarding ***privileged contexts and professional*** custody. Professions bound by strict confidentiality obligations—such as law, medicine, journalism, and finance—operate under established frameworks of privilege and secrecy. However, when practitioners in these fields utilize AI-mediated systems, they inadvertently introduce a third party into the privileged relationship, potentially compromising the very foundation of their professional duties.

The core issue is that AI platforms, by their nature, observe and process the information fed into them. When a lawyer uses an LLM to analyze a client's case file, or a physician uses an AI scribe to summarize a patient consultation, the AI system gains access to highly sensitive, privileged information. This access, even if intended solely for the practitioner's benefit, creates a custody event. The platform may retain logs, generate embeddings, or subject the data to automated or human review for safety or model improvement purposes. This opacity regarding data handling practices directly conflicts with the requirements of professional secrecy. If a platform cannot provide verifiable assurances—such as non-use receipts and transparent access provenance—that privileged data is isolated, protected, and immune from unauthorized access or secondary use, then the use of such AI tools may constitute a breach of professional ethics and legal obligations. The challenge for cognitive custody is to develop frameworks that allow professionals to leverage AI capabilities without compromising the sanctity of privileged contexts, ensuring that the introduction of a machine intermediary does not equate to a waiver of confidentiality.

Agents and Execution Risk

The emergence of AI agents—autonomous systems capable of executing tasks, interacting with tools, and making decisions on behalf of users—introduces a new and complex layer of ***execution risk*** within the domain of cognitive custody. Unlike traditional AI models that primarily process information, agents actively operate within a user's digital environment, potentially accessing, modifying, and creating data across various systems. This shift from passive processing to active execution fundamentally alters the custody landscape.

Execution risk arises from several factors: (1) ***Unforeseen Consequences***: Agents, particularly those with broad permissions and complex decision-making capabilities, can initiate actions that lead to unintended or undesirable outcomes. A seemingly innocuous instruction can cascade into a series of operations that compromise data, alter critical configurations, or expose sensitive information. (2) ***Permission Creep***: Over time, agents may accumulate or be granted expanded permissions that exceed their initial scope, often without explicit user awareness or re-consent. This silent privilege expansion creates vulnerabilities, as agents gain access to more sensitive areas of a user's cognitive and operational life-graph. (3) ***Lack of Audit-ability***: Tracing the precise sequence of decisions and actions taken by an autonomous agent can be challenging. The black-box nature of some agentic systems makes it difficult to ascertain *why* a particular action was taken, what data was accessed or modified, and *where* the resulting artifacts were routed. This opacity hinders accountability and complicates incident response.

Mitigating execution risk requires a robust framework for agent governance. This includes: (a) ***Signed Agent Manifests***: Requiring agents to declare their capabilities, permissions, and intended behaviors in a verifiable, cryptographically signed manifest. (b) ***Task-Scoped Execution***: Limiting agent actions to specific, narrowly defined tasks, with explicit user approval for each execution cycle. (c) ***Real-time Monitoring*** and Alerting: Implementing systems that provide users with real-time visibility into agent activities and alert them to any deviations from expected behavior or attempts to access sensitive data. (d) ***Rollback Capabilities***: Ensuring that users can revert agent-initiated changes or revoke permissions instantly. The governance of AI agents must evolve to provide users with verifiable control and transparency over the autonomous execution of tasks, transforming agentic trust into auditable execution provenance.

Open Source Does Not Equal Audited

The open-source movement has long been lauded for its transparency, collaborative development, and the promise of enhanced security through community scrutiny. The adage "given enough eyeballs, all bugs are shallow" reflects a core tenet of open-source security. However, in the context of AI-mediated systems, particularly those involving privileged agents and complex data pipelines, the notion that open source does not automatically equate to audited security becomes critically apparent. While the source code for many AI tools and

underlying infrastructure components may be publicly available, this visibility alone is insufficient to guarantee verifiable cognitive custody.

Several factors contribute to this disconnect: (1) Complexity and Interdependencies: Modern AI systems are often composed of numerous open-source libraries, frameworks, and models, each with its own dependencies and potential vulnerabilities. The sheer complexity of these interwoven components makes a comprehensive security audit a monumental task, often beyond the resources of individual users or even smaller organizations. (2) Dynamic Nature: Open-source projects are constantly evolving, with frequent updates, new features, and bug fixes. Maintaining a continuous audit of every change and its potential security implications is practically impossible for most users. (3) Focus on Functionality over Custody: Many open-source contributions prioritize functionality, performance, or novel AI capabilities, with less emphasis on the intricate custody implications of data access, retention, and processing. (4) Practical Audit-ability vs. Theoretical Transparency: While the code is theoretically auditable, the practical reality is that few users possess the expertise, time, or resources to conduct a thorough security review of every open-source component they utilize. Furthermore, even if a component is audited, its integration into a larger system can introduce new, unforeseen custody risks.

For AI agents and open-source LLM tooling, this gap is particularly salient. These tools often grant non-technical users significant technical reach, allowing them to interact with files, prompts, source code, credentials, calendars, and workflows. This means users become operationally technical before they are security-model technical. The availability of source code, therefore, should not be conflated with practical audit-ability or guaranteed custody. A robust cognitive custody framework must acknowledge this distinction and provide mechanisms for verifiable security beyond mere code transparency.

Fork Culture and Public R&D Extraction

Fork culture, a cornerstone of open-source development, is characterized by its productivity, creativity, and often, its elegance in fostering innovation. Independent builders frequently publish new features, workflow enhancements, user experience improvements, novel prompts, integrations, agent behaviors, patches, benchmarks, product ideas, tool wrappers, security fixes, and interface experiments. This vibrant ecosystem of shared innovation, however, inadvertently creates a significant asymmetry in the context of cognitive custody: the potential for public R&D extraction by larger entities.

While often legal under relevant open-source licenses, the ability of large companies to observe, absorb, and repackage these patterns with superior distribution channels, capital, legal capacity, and infrastructure raises profound ethical questions. Public repositories, designed for collaborative development, can effectively function as unpaid product intelligence pipelines. An open ecosystem, intended to democratize technology, can be leveraged as a comprehensive

market research platform, a mechanism for feature validation, a means of workflow discovery, and an implementation scaffolding for companies that did not originate the underlying insight. This is not an anti-open-source stance; rather, it is a custody aware position that highlights the need for mechanisms to ensure fair attribution and prevent the uncompensated extraction of intellectual labor. The challenge lies in balancing the benefits of open collaboration with the imperative to protect the cognitive contributions of individual creators from being silently absorbed and commercialized without verifiable provenance or equitable exchange.

Auto-Update and Moving Trust Boundaries

In the realm of agentic AI systems, the concept of trust is not a static, one-time decision made at the point of installation. Instead, it is a dynamic and continuously evolving relationship, profoundly impacted by the phenomenon of ***auto-update*** and the resulting ***moving trust boundaries***. A user may initially grant trust to version 1.0 of an AI agent based on its declared capabilities and permissions. However, subsequent versions—1.1, 1.2, and so forth—may introduce new telemetry, dependencies, endpoints, permissions, plugins, file access protocols, model providers, logging behaviors, auto-update mechanisms, Model Context Protocol (MCP) servers, or cloud routing configurations. Consequently, the original trust decision may no longer accurately reflect the expanded scope of the agent's operations.

This moving trust boundary problem is particularly critical for privileged agents that possess the ability to read files, invoke tools, and act across various systems. While auto-update offers convenience for ordinary software, silent privilege expansion in agentic systems constitutes a significant custody event. Such expansions should not occur without explicit disclosure, diffing (highlighting changes in permissions and capabilities), cryptographic signing, and user revocability. A robust agent runtime environment should therefore provide: (1) ***permission diffs*** that clearly articulate changes in access rights; (2) ***dependency diffs*** detailing new or altered external components; (3) ***network diffs*** indicating changes in communication endpoints; (4) ***telemetry diffs*** outlining modifications to data collection practices; (5) ***file-access diffs*** specifying new or expanded file system access; (6) ***signature verification*** to ensure the integrity and authenticity of updates; (7) ***rollback capabilities*** allowing users to revert to previous, trusted versions; and (8) a strict policy against silent privilege expansion. Without these mechanisms, auto-update transforms from a convenience feature into a continuous source of unmanaged custody risk, undermining user control and transparency. For example, an AI agent might auto-update overnight, silently expanding its permissions to include access to new file types or external services. A user who initially approved the agent for a specific task might find its capabilities, and thus its access to their cognitive and operational life-graph, significantly broadened without explicit re-consent or even notification. This silent privilege expansion creates a continuous, unmanaged custody risk, undermining user control and transparency.

Data as the New Value Layer

In AI-mediated systems, the traditional understanding of value, often centered on the tool itself, is being fundamentally reshaped. While the AI tool provides the interface and functionality, the true asset, at scale, is increasingly the ***cognition itself, manifested as data.*** This paradigm shift means that the tool may not be the primary source of value; rather, it is the vast quantities of data generated through user interaction that become the new, highly valuable layer.

AI tools possess an unparalleled capacity to observe and learn from human activity. They can discern: (1) what individuals are building; (2) the specific challenges they encounter; (3) the features they actively seek; (4) the prompts that yield effective results; (5) the workflows that are frequently repeated; (6) the files they connect and integrate; (7) the errors that commonly occur; (8) the inventions that emerge; (9) the code patterns that recur; (10) the industries that are actively experimenting; (11) the products that are in formation; (12) the naming systems that appear; and (13) the clustering of user intentions. This comprehensive observational capability creates a profound ***extraction risk***, even in the absence of direct content copying.

A platform can learn from the aggregate patterns of user behavior, demand, failure, and innovation. It can identify which features are most needed, which workflows are most effective, which product categories are gaining traction, and where integration gaps exist that users are attempting to bridge. The pattern itself, derived from the collective cognitive and operational efforts of its users, becomes an invaluable asset. This necessitates a re evaluation of custody, extending it beyond individual data points to encompass the emergent value layer created by aggregated cognitive data. Without clear provenance and control over this new value layer, users risk contributing their intellectual capital to platforms without equitable recognition or verifiable oversight.

Cognitive Fingerprinting and Cross-Account Identity Reconstruction

In the digital age, identity reconstruction extends far beyond conventional identifiers such as names, email addresses, or device fingerprints. Humans possess unique ***creative and cognitive signatures*** that, when analyzed by AI systems, can facilitate ***cross-account identity reconstruction.*** These signatures are embedded in a multitude of digital behaviors and outputs, making individuals linkable even across ostensibly separate accounts or workspaces.

Such cognitive fingerprints can include, but are not limited to: naming conventions, sentence rhythm, preferred metaphors, file structures, repository organization, prompt style, design aesthetics, code style, commit habits, time-of-day activity patterns, tool choices, recurring project vocabulary, brand semantics, document formatting, calendar patterns, contact overlap, payment behaviors, domain registrations, uploaded files, and shared collaborators. Each of these elements

contributes to a unique digital persona that, when aggregated and analyzed, can reveal continuity across disparate online presences.

The analogy to music is apt: just as some producers are recognizable by their distinct sound selection, count-in, rhythm, texture, and arrangement before their name appears in the credits, individuals can leave a similar imprint on their cognitive outputs. A person may intentionally create different accounts, utilize distinct email addresses, and maintain separate workspaces to compartmentalize their digital life. However, if their underlying cognitive patterns remain consistent, AI systems can infer continuity and link these seemingly disconnected identities. This means that account separation, a common strategy for privacy and compartmentalization, fails when cognition itself becomes linkable. The challenge for cognitive custody, therefore, is to provide mechanisms that prevent or at least make transparent such cross-account identity reconstruction, ensuring that users retain control over their fragmented digital selves. For instance, a creator using multiple accounts across different platforms might still be linkable through consistent naming conventions, unique file structures, distinctive writing styles, or characteristic code styles. AI systems, by analyzing these subtle cognitive fingerprints, can reconstruct a unified identity even when the user has intentionally fragmented their digital presence, thereby undermining efforts at pseudonymity or compartmentalization.

Linkage Transparency and the Right to Unlink

Given the advanced capabilities of AI platforms to link accounts, identities, devices, documents, writing patterns, domains, payment systems, connectors, and behavioral signals, users must be afforded both ***linkage transparency*** and the ***right to unlink.*** While such linkages may be necessary for legitimate purposes—such as fraud prevention, abuse mitigation, security protocols, billing accuracy, or maintaining platform integrity—they should not operate invisibly or without user recourse when they impact an individual's identity graph.

A comprehensive ***linkage provenance record*** is proposed as a foundational element of cognitive custody. Such a record would meticulously document each linkage event, including: (1) a unique linkage event ID; (2) the hash of the source account; (3) the hash of the target account; (4) the explicit basis for the linkage (e.g., shared identifiers, behavioral signals, document overlap, device or network signals, connector overlap); (5) whether the linkage was human-reviewed or automated; (6) a confidence score associated with the linkage; (7) the stated purpose of the linkage; (8) the retention policy for the linkage record; (9) user visibility status; (10) an appeal route for contesting the linkage; (11) a timestamp of the event; and (12) a cryptographic signature for integrity verification. If AI systems are capable of inferring continuity across disparate digital footprints, users must be granted equivalent visibility into these inferences. The right to unlink, supported by such transparent provenance, empowers individuals to challenge and, where appropriate, sever connections that are inaccurate, unwarranted, or no longer desired, thereby restoring agency over their digital identity. For example, if an AI system correlates a user's professional identity (e.g., GitHub contributions, LinkedIn profile) with their personal

identity (e.g., social media posts, personal blog) through stylometric analysis or shared metadata, the user should have the right to transparently view these linkages and request their dissolution. This ensures that the user, not the AI, ultimately defines the boundaries and connections of their own identity.

From Data Breach to Life-Graph Breach

The traditional understanding of a ***data breach*** typically involves the unauthorized access or exposure of specific records, such as passwords, credit card numbers, or personal identifying information. While serious, these breaches often represent a compromise of discrete data points. However, in the era of AI-mediated systems, the interconnectedness of cognitive and operational data elevates the risk to a new and far more comprehensive level: the life-graph breach. A ***life-graph breach*** is not merely the leakage of records; it is the exposure of the underlying structure, patterns, and predictive insights that constitute an individual's entire digital existence.

Consider the distinction: a password breach might expose an account credential, allowing unauthorized access to a single service. A life-graph breach, by contrast, can expose a multitude of interconnected elements: aliases used across different platforms, unique writing styles, confidential company plans, intricate calendar patterns, a network of collaborators, pending business deals, the structure of personal files, financial pressures, family routines, and even precise future locations. AI-mediated systems significantly amplify the risk of such breaches because they are designed to collect, aggregate, and connect heterogeneous contexts—thoughts, files, schedules, relationships, operational patterns, identity markers, and future intentions. If such a system, or any component within its supply chain (e.g., a vendor, connector, logging layer, support queue, or derived-data store), is compromised, the attacker gains not just isolated data, but a comprehensive, actionable map of an individual's life. This transformation from a data breach to a life-graph breach represents a critical escalation in the potential harm and necessitates a re-evaluation of security protocols and custodial responsibilities.

Adversarial Reconstruction: When Leaked AI Context Becomes a Targeting Map

For individuals of high net worth, founders, researchers, executives, celebrities, political figures, journalists, lawyers, activists, inventors, and their families, the most significant and insidious risk stemming from AI-mediated systems is not merely embarrassment or financial loss, but ***adversarial reconstruction.*** This refers to the process by which a malicious actor utilizes leaked AI-mediated context to meticulously map vulnerabilities, predict future locations, identify critical relationships, ascertain financial pressures, uncover business dependencies, exploit legal leverage, understand personal routines, and ultimately devise strategies to harm or coerce a person or company.

An adversary, armed with such comprehensively reconstructed intelligence, can gain insights into: an individual's future whereabouts; the identities of those they trust; their ongoing projects and intellectual property; their deepest fears and anxieties; pending business deals and negotiations; existing lawsuits or legal challenges; critical accounts and digital assets; vendors and partners they rely upon; the assistants or employees who have privileged access; predictable routines and habits; weaknesses in their security practices; family relationships that could be leveraged; identity aliases that link back to them; and projects that are not yet public. The highest-risk breach, therefore, is not the exposure of a single file or a discrete piece of information. It is the holistic reconstruction of the person or institution behind many files, transforming seemingly disparate data points into a potent targeting map. This capability fundamentally shifts the landscape of digital security, demanding a proactive and comprehensive approach to cognitive custody that anticipates and mitigates the risks of such sophisticated adversarial exploitation.

Privileged Contexts and Professional Custody

Certain cognitive chains are not merely private; they are imbued with legal, professional, or ethical privilege, necessitating an even higher standard of custody. Professionals such as lawyers, doctors, therapists, journalists, accountants, researchers, executives, patent inventors, fiduciaries, and regulated professionals routinely process context that carries duties extending far beyond ordinary privacy expectations. The integration of AI-mediated workflows into these domains introduces significant complexities and potential compromises to these established privileges.

A legal memo, a therapeutic note, a medical record, a source interview, unpublished research findings, an audit file, sensitive board communications, a patent disclosure, or a fiduciary analysis cannot be treated as ordinary productivity context. These materials are subject to stringent ethical codes, legal protections, and professional responsibilities. Therefore, ***professional custody*** in AI-mediated systems demands elevated tiers of protection, including:

- ***Stronger Access Logs:*** Granular and immutable records of all access events, detailing who, what, when, and why.

- ***Stricter Retention Policies***: Data retention periods must align with professional and legal obligations, with clear deletion protocols.

- ***No Default Training Use***: Privileged data must be explicitly excluded from model training or product improvement processes unless specific, informed consent is obtained.

- ***Explicit Human-Review Disclosure***: Any instance of human review of privileged data must be clearly disclosed to the user, with a verifiable audit trail.

- ***Privilege-Preserving Workflows***: AI tools must be designed to maintain and respect professional privileges, potentially through anonymization, pseudonymization, or secure enclaves.

- ***Connector Restrictions***: Strict controls on third-party connectors to prevent unauthorized data egress or entanglement with less secure systems.

- ***Local-First Options***: Providing capabilities for processing highly sensitive information locally, minimizing cloud exposure.

- ***Exportable Evidence of Custody Boundaries***: Mechanisms for users to demonstrate that appropriate custody measures were in place, particularly crucial for compliance and legal defense.

AI-mediated workflows, while offering efficiency gains, must be meticulously designed to uphold these professional duties. Failure to do so risks not only legal and ethical breaches but also a fundamental erosion of trust in professions that rely on the sanctity of privileged information.

Cognitive Supply-Chain Risk

The concept of cognitive supply-chain risk emerges as a critical counterpart to the well-established software supply-chain risk. While software supply-chain risk primarily interrogates what code enters a system, cognitive supply-chain risk asks a more profound question: what human thought, in its nascent and evolving forms, leaves it? This risk refers to the potential for cognition-in-progress to traverse a complex chain of AI systems, tools, vendors, logs, models, connectors, agents, and identities without verifiable custody, transparent access provenance, or clear downstream-use accountability.

This new vector of risk encompasses a broad array of potential harms, including:

- ***Intellectual Property (IP) Leakage***: Unauthorized disclosure of proprietary ideas, designs, or creative works during their formation.

- ***Trade-Secret Exposure***: Compromise of confidential business information, strategies, or methodologies.

- ***Credential Exposure***: Unintended revelation of access credentials or sensitive authentication tokens.

- ***Unconsented Review***: Human or automated review of cognitive artifacts without explicit user permission.

- ***Hidden Retention***: Storage of cognitive data beyond stated retention policies or without user knowledge.

- ***Model-Improvement Use***: Utilization of user-generated cognitive data for training or refining AI models without consent.

- ***Derivative Product Learning:*** Platforms gaining insights from aggregated user cognition to develop competing products or features.

- ***Connector Leakage:*** Data egress through third-party connectors that lack adequate custody controls.

- ***Agent Exfiltration***: Autonomous agents inadvertently or maliciously transferring sensitive cognitive data.

- ***Open-Source Dependency Compromise***: Vulnerabilities in open-source components leading to cognitive data exposure.

- ***Safety-Review Overreach***: Access to cognitive data under the guise of safety review that exceeds its legitimate scope.

- ***Legal Disclosure Without Timely Notice***: Revelation of cognitive data in legal contexts without appropriate user notification.

- ***Cross-Account Identity Reconstruction***: Linking of disparate user identities through cognitive patterns.

- ***Life-Graph Breach***: Comprehensive exposure of an individual's entire operational and cognitive existence.

- ***Adversarial Targeting***: Use of leaked cognitive context to map vulnerabilities and target individuals or organizations.

Effectively managing cognitive supply-chain risk requires a holistic approach that extends custody principles across the entire ecosystem of AI-mediated interactions, ensuring that every link in the chain is auditable and accountable for the human thought it processes.

Access Provenance as a Missing Primitive

The next generation of AI governance must fundamentally shift its focus to track not only the origin and authenticity of content but, more critically, who had custody of the cognition before that content ever existed. While existing provenance work has made significant strides in addressing ***content provenance***—asking questions such as who made an output, whether it was AI-generated, whether it was altered, and where it came from—this paper argues for a broader, upstream primitive: ***access provenance***.

Content provenance initiatives, such as Content Credentials and C2PA, represent major progress in verifying the authenticity and origin of digital media. Similarly, systems like SynthID and other watermarking technologies are advancing the identification of AI-generated outputs. Research into systems like Tree-Ring Watermarking further demonstrates the potential to influence generative processes rather than merely modifying outputs post-generation. These advancements are undoubtedly crucial. However, they address only one part of the custody problem.

Access provenance, in contrast, delves into the foundational questions of: who accessed an artifact; why that access occurred; when it happened; by what authority it was granted; what was the scope of the access; through which system it was facilitated; whether the accessed information was retained; whether it was transformed; whether it was disclosed; and whether it was linked to other identities or systems. This is a missing primitive in the current governance landscape. If AI systems are actively participating in the formation of intellectual work, then provenance must extend beyond mere outputs to encompass access events, transformation chains, execution paths, identity linkages, and the precise boundaries of custody. Without robust access provenance, the ability to audit and verify the integrity of the cognitive supply chain remains fundamentally compromised.

Proposed Framework: Cognitive Custody Receipts

To address the pervasive issues of unverifiable trust and opaque data handling in AI-mediated systems, this paper proposes a receipt-based governance framework. This framework shifts the burden of proof from implicit trust to explicit, auditable evidence of custody. A receipt, in this context, does not necessitate the public exposure of sensitive content; rather, it serves as a verifiable record of critical custody facts. These facts include: the class of artifact accessed, the authority under which access occurred, the scope of that access, the specific system involved, the purpose of the access, the retention policy applied, the downstream route of the data, and a cryptographic signature to ensure integrity.

This comprehensive framework includes several distinct types of receipts, each designed to provide granular transparency across the cognitive supply chain:

- ***Access Receipts***: Document individual access events to any cognitive or operational artifact, detailing who, what, when, why, and how.

- ***Cognitive Chain Receipts***: Record the formation pathway of an idea, from initial prompt to final artifact, including all AI interactions and transformations.

- ***Operational Access Receipts***: Provide verifiable records for connector activity, detailing data flow and access across integrated third-party services.

- ***Agent Execution Receipts***: Document the actions and decisions made by autonomous AI agents, including tool invocations and data modifications.

- ***Connector Custody Manifests***: Offer a comprehensive overview of data handling policies and access permissions for each integrated connector.

- ***Disclosure Boundary Tags***: Mark sensitive data with explicit classifications that dictate permissible disclosure and usage.

- ***Identity-Linkage Provenance Records***: Document any correlation or linking of disparate user identities by AI systems, including the basis and confidence score.

- ***Training/Non-Use Receipts***: Provide verifiable attestations from platforms regarding whether user data has been used for model training or other specified purposes.

By converting abstract trust claims into inspectable custody evidence, this receipt-based framework empowers users with the necessary tools to audit and verify the custodial practices of AI platforms. It transforms the governance of thought, files, identity, families, and operational life-graphs from an opaque reliance on policy statements into a transparent, evidence-driven system of accountability.

Receipt Types

Receipt Type	***What It Records***	***When Generated***	***Who Can Audit It***
Cognitive Chain Receipt	The evolution of ideas and intellectual work through AI-mediated processes, including prompts, drafts, and revisions.	Upon each significant interaction or save event within the cognitive chain.	User, designated auditor, platform (with user consent).
Operational Access Receipt	All AI-mediated operational access to an individual's life- graph, including schedules, communications, and files.	Upon each instance of AI access to operational data.	User, designated auditor, platform (with user consent).
Vector Shadow Provenance Record	Creation, storage, and utilization of derived data (embeddings, summaries) from user inputs.	Upon creation or modification of vector shadows.	User, designated auditor, platform (with user consent).
Identity-Linkage Provenance Record	Instances where AI systems correlate or link disparate user identities across accounts or behavioral patterns.	Upon establishment or modification of identity linkages.	User, designated auditor, platform (with user consent).
Agent Execution Receipt	Actions and decisions made by autonomous AI agents, including tool invocations and data modifications.	Upon each agent action or decision.	User, designated auditor, platform (with user consent).
Connector Custody Manifest	Data handling policies and access permissions for each integrated third-party connector.	Upon connector integration or policy update.	User, enterprise administrator, platform, regulatory bodies.
Non-Use Receipt	Verifiable attestation that user data has not been used for specified purposes (e.g., model training).	Upon request or periodically, as per user settings.	User, designated auditor, regulatory bodies.
Disclosure Boundary Tag	Metadata indicating the sensitivity and intended disclosure scope of specific data elements.	Upon data creation or classification.	User, platform, relevant stakeholders.

User-Facing Access Ledger

The user-facing access ledger serves as the interface layer of cognitive custody, translating complex backend data interactions into an understandable and auditable record for the individual. Every AI platform handling sensitive cognitive or operational context should be mandated to provide such a ledger, offering users a transparent window into how their data is being managed. This ledger should detail, for each access event:

- ***Accessed By***: The specific AI system, agent, or human entity that initiated the access.
- ***Reason***: The explicit justification or purpose for the access (e.g., user-initiated request, automated workflow, safety review, support inquiry, legal hold).
- ***Time***: The precise timestamp of the access event.
- ***Scope***: The exact data points, files, conversations, or system components that were accessed.
- ***System***: The particular AI model, tool, or service involved in the access.
- ***Human-or-Automated Status***: A clear distinction indicating whether the access was performed by a human operator or an automated process.
- ***Files Touched***: A list of specific files or documents that were interacted with.
- ***Conversation Touched***: Identification of any conversational data that was accessed.
- ***Connectors Used***: Details of any third-party connectors involved in the data flow.
- ***Data Retained***: Information on whether the accessed data was retained, and if so, for what duration.
- ***Data Disclosed***: Indication of whether the data was disclosed to any external entities.
- ***Retention Period***: The specified period for which the data will be stored.
- ***Deletion Status***: Confirmation of data deletion or scheduled deletion.
- ***Identity Linkage***: Disclosure of any instances where the accessed data was linked to other identities or accounts.
- ***Third-Party Route***: Information on any external routing or processing of the data.

- ***Revocation Status***: The current status of user permissions and the ability to revoke access.

The ledger must be designed to be accessible and comprehensible to non-technical users, yet sufficiently detailed to meet the scrutiny of professionals, enterprises, auditors, and regulators. It should clearly differentiate between various types of access, such as user-requested actions, automated retrievals, safety reviews, support interventions, enterprise administrative access, legal holds, third-party connector access, model-provider processing, vector indexing, memory utilization, and agent execution. By providing this level of granular transparency, the user-facing access ledger transforms the abstract concept of custody into a concrete, verifiable reality, empowering individuals to exercise genuine control over their digital lives.

Enterprise Implications

For enterprises, the challenges of cognitive custody escalate from individual concerns to institutional vulnerabilities. The widespread adoption of AI without robust custody controls transforms potential individual data leaks into systemic risks, exposing critical organizational assets. In this context, ***cognitive custody*** becomes ***institutional custody***, and its absence can lead to severe consequences, including the compromise of trade secrets, customer data, legal privilege, employee records, codebases, product roadmaps, incident reports, board materials, HR records, contracts, financial forecasts, and the very institutional memory of the organization. The proliferation of "shadow AI"—unauthorized or unmanaged AI tool usage within an enterprise—and the sprawl of interconnected data via numerous connectors further exacerbate this problem, turning governance into a complex graph problem where data flows are difficult to track and control.

To mitigate these profound risks, enterprises require a comprehensive suite of AI custody measures:

- ***AI Access Logs***: Granular logging of all AI interactions with enterprise data, detailing access events, users, systems, and purposes.

- ***Agent Permission-ing***: Strict, centrally managed permissioning frameworks for AI agents, ensuring they operate only within authorized boundaries.

- ***Connector Inventories***: A complete and continuously updated inventory of all third-party connectors, along with their data access permissions and security postures.

- ***Local-First Sensitive Workflows***: Implementation of workflows that prioritize local processing for highly sensitive data, minimizing cloud exposure.

- ***Signed Tool Manifests***: Requiring AI tools and agents to provide verifiable, cryptographically signed manifests of their capabilities and data handling practices.

- ***Retention Controls***: Enterprise-wide policies and technical controls for data retention and deletion across all AI-mediated systems.

- ***Admin Transparency***: Enhanced transparency for enterprise administrators regarding their access to user and organizational data within AI platforms.

- ***Employee Training***: Comprehensive training programs for employees on AI custody best practices, risk awareness, and responsible AI usage.

- ***Procurement Standards for AI Custody***: Integrating robust custody requirements into the procurement process for all AI tools and services.

- ***Task-Scoped Consent***: Implementing mechanisms for granting AI access based on specific tasks, rather than broad permissions.

- ***Identity-Linkage Transparency***: Providing visibility into how AI systems link employee identities across different platforms and data sources.

- ***Breach Compartmentalization***: Designing systems and workflows to limit the scope and impact of potential data breaches.

- ***Operational Access Receipts***: Generating verifiable receipts for all AI-mediated operational access to enterprise systems and data.

Without these critical safeguards, enterprise AI adoption risks becoming a conduit for institutional memory leakage at scale, undermining competitive advantage, regulatory compliance, and stakeholder trust. Effective cognitive and operational custody is therefore not merely a technical requirement but a strategic imperative for modern enterprises.

Individual Creator and Inventor Implications

For independent builders, artists, researchers, founders, writers, and inventors, the implications of AI-mediated cognition extend beyond general privacy concerns to encompass significant personal and strategic risks. These individuals, often operating without the institutional safeguards available to larger enterprises, may inadvertently expose highly sensitive pre-filing patent ideas, trade secrets, nascent naming systems, evolving product strategies, foundational code architectures, creative directions, personal notes, financial constraints, and even emotional motivations through their AI-mediated workflows, long before formal protections can be established.

The independent creator in the AI era thus requires a new form of ***lab-notebook discipline*** for their AI-mediated thought processes. This discipline necessitates:

- ***Local-First Workflows***: Prioritizing local processing and storage for the most sensitive material, minimizing exposure to cloud-based AI systems.

- ***Timestamped Invention Records***: Maintaining verifiable, timestamped records of ideas and developments, crucial for establishing intellectual property ownership.

- ***Disclosure Boundaries***: Clearly defining and managing the boundaries of what information is shared with AI systems and under what conditions.

- ***Hashed Drafts***: Utilizing cryptographic hashing to create immutable records of drafts at various stages, providing proof of prior art or creative evolution.

- ***Controlled AI Usage***: Exercising deliberate and informed control over which AI tools are used for specific tasks, understanding their data handling policies.

- ***Receipts for Model and Tool Exposure***: Demanding and retaining verifiable receipts that document when and how cognitive data was exposed to AI models and tools.

- ***Strict Connector Hygiene***: Meticulously managing and auditing third-party connectors to prevent unintended data leakage.

- ***Separation of Sensitive Workspaces***: Maintaining distinct digital environments for highly sensitive projects, isolated from general-purpose AI interactions.

- ***Account-Linkage Awareness***: Understanding how AI systems might link disparate accounts or identities and taking steps to mitigate unwanted correlations.

This does not imply a complete abstention from AI tools, which offer immense creative and productive potential. Rather, it advocates for using AI with a profound ***custody awareness***, ensuring that the benefits of AI augmentation do not come at the cost of intellectual sovereignty and personal strategic advantage.

Regulatory Recommendations

To effectively address the challenges posed by AI-mediated cognition and operational custody, regulatory bodies and standards organizations must evolve their approach from focusing solely on privacy promises to demanding verifiable custody evidence. This necessitates a proactive and comprehensive regulatory framework that mandates transparency and accountability across the AI ecosystem. Key regulatory recommendations include:

Mandatory Access Logs for AI Platforms: Requiring all AI platforms to maintain granular, immutable access logs for every interaction with user data, detailing the who, what, when, why, and how of each access event.

User-Visible Human-Review Disclosure: Mandating clear and timely disclosure to users whenever their data is subjected to human review, where legally permissible, distinguishing between automated and human interventions.

Reason Codes for Access: Requiring platforms to provide explicit reason codes for any access to user data, ensuring transparency regarding the justification for such access.

Scope Disclosure: Enforcing clear disclosure of the precise scope of data accessed during any interaction, preventing broad, undefined data collection.

Connector Access Transparency: Implementing regulations that demand full transparency regarding data flows and access permissions for all third-party connectors integrated with AI platforms.

Retention and Deletion Receipts: Requiring platforms to provide verifiable receipts for data retention periods and confirmation of data deletion, ensuring compliance with user preferences and legal obligations.

Training-Use Receipts: Mandating explicit receipts or attestations from platforms regarding whether user data has been utilized for model training or other product improvement purposes.

Legal-Hold Notice: Requiring timely notification to users when their data is subject to a legal hold, where permissible.

Enterprise Admin Access Logs: Implementing robust logging and audit trails for enterprise administrators accessing user data within AI platforms.

Signed Agent Manifests: Requiring AI agents to provide cryptographically signed manifests detailing their capabilities, permissions, and intended behaviors.

No Silent Privilege Expansion: Prohibiting AI systems from silently expanding their access privileges or capabilities without explicit user consent and verifiable disclosure.

Independent Audits of AI Custody Systems: Mandating regular, independent audits of AI platforms' custody systems to verify compliance with regulatory requirements.

Special Protections for Privileged Contexts: Establishing enhanced protections for patent-pending, legal, medical, financial, child-related, family-related, and trade-secret contexts, recognizing their heightened sensitivity.

Identity-Linkage Transparency: Requiring platforms to disclose when and how they link disparate user identities, providing users with the right to contest and unlink.

Task-Scoped Consent Requirements: Promoting and, where appropriate, mandating task-scoped consent models that limit AI access to specific, narrowly defined purposes.

Operational Access Receipts for Connectors: Extending the receipt framework to cover all operational access events facilitated by third-party connectors.

Breach Notification Standards for Life-Graph Exposure: Developing specific breach notification standards tailored to the comprehensive nature of life-graph breaches, ensuring users are adequately informed of the scope of compromise.

The overarching aim of these regulatory recommendations is not to stifle innovation or impede useful workflows, but rather to ensure that AI systems, which increasingly demand and rely upon user trust, are capable of producing verifiable evidence of their custodial practices. This aligns with emerging frameworks like the EU AI Act, which establishes comprehensive rules on AI risk and transparency [10]. Furthermore, with AI-related incidents jumping significantly in recent years [11], verifiable custody is a necessity. This shift is essential for fostering a secure, transparent, and accountable AI ecosystem.

Intus Provenance Infrastructure

The implications for ***provenance infrastructure***, particularly for entities like INTUS, are direct and transformative. Existing provenance work has largely concentrated on outputs: verifying the origin and authenticity of content, determining if an image was AI-generated, whether media was altered, who created a file, or if content is authentic. These are undeniably necessary questions for establishing trust in digital artifacts. However, the cognitive custody problem fundamentally begins upstream. If AI systems are now deeply embedded in the formation of intellectual work, then provenance must extend beyond mere outputs to encompass access events, transformation chains, execution paths, identity linkages, and the precise boundaries of custody.

This necessitates a conceptual expansion towards a ***composable cognitive custody infrastructure***. This infrastructure would integrate various layers of provenance:

- ***Content Provenance***: Verifying the origin and integrity of finished artifacts.
- ***Access Provenance***: Documenting who accessed what, when, why, and how within AI systems.

- ***Cognitive-Chain Provenance***: Tracking the evolution of ideas and intellectual work through AI-mediated processes.

- ***Operational Custody Provenance***: Recording interactions with an individual's operational life-graph via AI tools and connectors.

- ***Execution Provenance***: Documenting the actions and decisions of AI agents.

- ***Identity-Linkage Provenance***: Providing transparency and audit-ability for cross-account identity correlations.

INTUS, as a substrate-level provenance infrastructure, is uniquely positioned within this landscape. Its role would not be merely to append a label after an output has been generated, but to function as a foundational receipt layer that records custody from the very moment work enters its formation stage. The strongest claim is not that no one is working on generation-time watermarking [3, 20] or content credentials [1, 2]; rather, it is that the market currently lacks a general, comprehensive custody layer for cognition, operations, connectors, agents, and access events. By providing this foundational layer, INTUS can enable the verifiable trust and accountability essential for the future of AI-mediated intellectual work.

The Right to Know Who Touched the Thought, the File, The Identity, the Family Orbit, and the Life-Graph

In an increasingly AI-mediated world, the remedy for paranoia is not reassurance; the remedy is receipts. AI systems have entered the workshop, the lab notebook, the studio, the codebase, the calendar, the family home, and the very formation layer of human cognition and operational life. This pervasive integration, while offering unprecedented efficiencies and capabilities, has simultaneously created a profound and systemic custody problem. Users, whether individuals, creators, or enterprises, are increasingly operating within an architecture of invisible trust, where their most sensitive thoughts, files, identities, family orbits, and life-graphs are accessible to systems whose internal operations remain largely opaque.

This paper has argued that traditional frameworks for privacy and intellectual property are insufficient to address the unique challenges posed by AI-mediated cognition. The composition problem, the asymmetry of visibility, the emergence of derived data and vector shadows, the non-use problem, the entanglement of data through connectors, and the profound risks of life-graph breaches and adversarial reconstruction all underscore the urgent need for a new paradigm. We have introduced the concepts of ***cognitive custody*** and ***operational custody*** to precisely name these challenges, emphasizing the critical demand for ***access provenance***.

The proposed framework of Cognitive Custody Receipts—encompassing access receipts, cognitive chain receipts, operational access receipts, agent execution receipts, connector custody

manifests, disclosure boundary tags, identity-linkage provenance records, and training non-use receipts—offers a concrete, verifiable solution. These receipts, presented through a user-facing access ledger, transform abstract policy promises into auditable evidence, empowering users with the right to know: who touched their thought, their file, their identity, their family orbit, and their life-graph; why that access occurred; what was touched; and where those artifacts went. This shift from implicit trust to explicit, verifiable accountability is not merely a technical requirement; it is a fundamental imperative for preserving human agency, intellectual sovereignty, and personal security in the AI era. By embracing this receipt-based governance, we can build an AI ecosystem where trust is not assumed but demonstrably earned, ensuring that the transformative power of AI is harnessed responsibly and ethically.

Key Terms and Definitions

- ***Cognitive Custody***: The condition in which a third-party system has access to, influence over, or operational control around cognition-in-progress.

- ***Operational Custody***: Third-party access to the connected graph of real-world activity: schedules, communications, files, finances, contacts, tasks, locations, code, obligations, dependents, and decision flows.

- ***Cognition-in-progress***: Prompts, drafts, files, source code, invention trails, research notes, product strategy, naming systems, legal reasoning, business logic, emotional reasoning, uploaded documents, connected data, agent logs, memory artifacts, embeddings, summaries, derived metadata, revision pathways, and intermediate outputs.

- ***Cognitive Chain***: The pre-publication pathway through which a thought becomes an artifact in AI-mediated systems, including raw thought, prompt, model interaction, uploaded file, retrieved context, memory layer, draft, revision, code generation, agent execution, connector access, cloud routing, logs, telemetry, final artifact, publication, filing, or deployment.

- ***Life-Graph***: The comprehensive, interconnected web of an individual's real-world activities, relationships, and intentions, as reconstructed and inferred by AI-mediated systems.

- ***Family Orbit***: The network of individuals and their associated data (e.g., schedules, locations, communications) that are directly or indirectly exposed through a user's interaction with AI systems, particularly concerning dependents and household members.

- ***Access Provenance***: The right to know who touched a thought, file, identity, family orbit, or operational graph; why access occurred; what was touched; and where those artifacts went.

- ***Content Provenance***: The origin and history of digital content, including who made it, whether it was AI-generated, whether it was altered, and where it came from.

- ***Vector Shadow***: Computationally generated artifacts (e.g., embeddings, summaries) that capture semantic meaning, relationships, and patterns from a user's cognition-in-progress and operational life-graphs, often persisting even after original data deletion.

- ***Non-Use Receipt***: A verifiable attestation from a platform confirming that user data has not been used for specified purposes, such as model training or product improvement.

- ***Connector Custody Manifest***: A comprehensive overview of data handling policies and access permissions for each integrated third-party connector in an AI system.

- ***Agent Execution Receipt***: A verifiable record documenting the actions and decisions made by autonomous AI agents, including tool invocations and data modifications.

- ***Identity-Linkage Provenance***: Transparency and audit-ability for how AI systems correlate or link disparate user identities across accounts, devices, or behavioral patterns.

- ***Task-Scoped Consent***: Granular, context-specific authorization for AI system access to data, explicitly tied to a particular task or workflow, defining what data can be accessed, for what purpose, and for what duration.

- ***Custody Tiers***: A framework for categorizing data based on its inherent sensitivity and the potential impact of its compromise, defining proportionate security measures and disclosure protocols.

References

C2PA / Content Credentials
Coalition for Content Provenance and Authenticity (C2PA). (2024). C2PA Technical Specification v2.4. Available at: https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html

Content Credentials
Content Credentials. (n.d.). Available at: https://contentcredentials.org/

SynthID (Text Watermarking)
Dathathri, S., et al. (Google DeepMind). (2024). Scalable watermarking for identifying large language model outputs. Nature.
DOI: 10.1038/s41586-024-08025-4
Available at: https://www.nature.com/articles/s41586-024-08025-4

AI Agent Security / Prompt Injection
Deng, Z., et al. (2025). AI agents under threat: A survey of key security challenges and future pathways. ACM Computing Surveys.
DOI: 10.1145/3716628
Available at: https://dl.acm.org/doi/abs/10.1145/3716628

Indirect Prompt Injection in Agents
Dziemian, M., et al. (2026). How Vulnerable Are AI Agents to Indirect Prompt Injections? Insights from a Large-Scale Public Competition. arXiv preprint arXiv:2603.15714.
Available at: https://arxiv.org/abs/2603.15714

Embedding Inversion Attacks
Qin, S., et al. (2025). Uncovering Risks of Data-Free Feature Vector Inversion Attacks Against Vector Databases. IEEE Transactions on Information Forensics and Security.
DOI: 10.1109/TIFS.2025.11146802
Available at: https://ieeexplore.ieee.org/abstract/document/11146802/

Code Stylometry / De-anonymization
Caliskan-Islam, A., et al. (2015). De-anonymizing Programmers via Code Stylometry. 24th USENIX Security Symposium.
Available at: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/caliskan-islam

AI Privacy Risks for Children
McAfee Online Safety Research. (2025). Cyberbullying and AI Manipulated Content Are Top Parents' Concerns. McAfee Newsroom.
Available at: https://www.mcafee.com/en-us/newsroom/press-releases/2025/cyberbullying-and-ai-manipulated-content-top-parents-concerns-mcafee-kids-online-safety-research-finds.html

Legal Professional Privilege and AI
WilmerHale Client Alert. (2026, March 20). Protecting Legal Professional Privilege in the Age of AI. WilmerHale.
Available at: https://www.wilmerhale.com/en/insights/client-alerts/20260320-protecting-legal-professional-privilege-in-the-age-of-ai

EU AI Act (Regulation 2024/1689)
European Parliament and Council. (2024, June 13). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence.
Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

AI Incident Statistics
Stanford University. (2025, April). 2025 AI Index Report.
Available at: https://aiindex.mit.edu/report/

Tree-Ring Watermarking (Diffusion Images)
Wen, Y., et al. (2023/2024). Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. NeurIPS 2023 / ECCV 2024 (RingID).
Available at: https://arxiv.org/abs/2305.20030

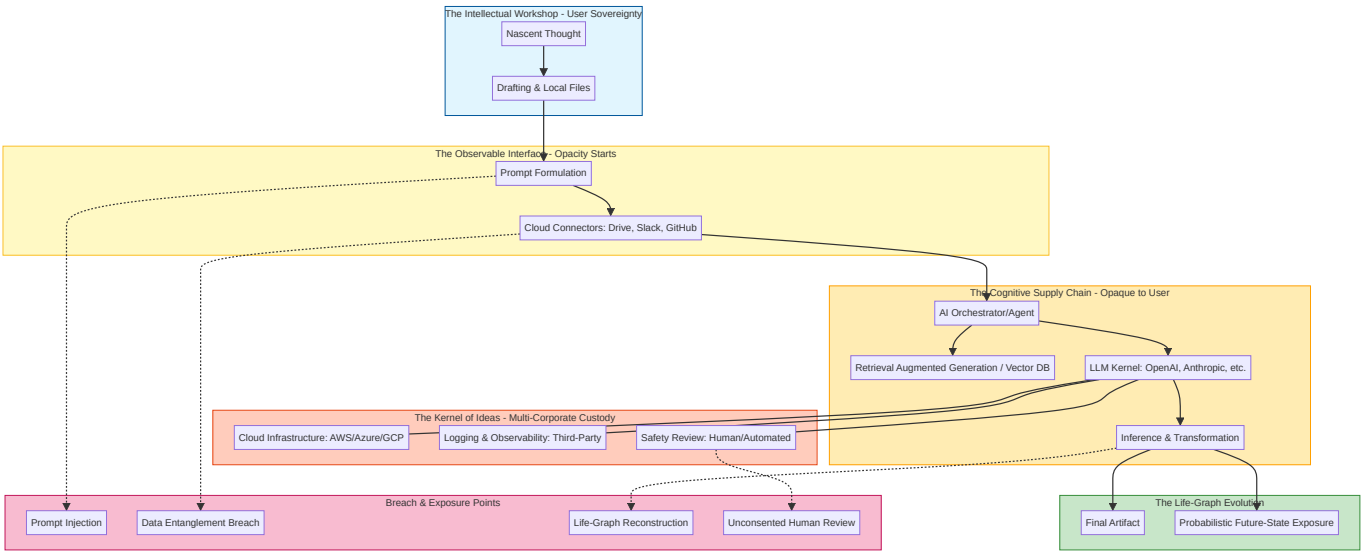
Cognitive Custody in AI-Mediated Systems: Information Flow, Opacity, and Risks

Executive Summary

This report analyzes the intricate flow of information within Large Language Model (LLM) mediated systems, drawing insights from the paper “Cognitive Custody: Access Provenance, Asymmetric Trust, and the Governance of Thought, Files, Identity, Families, and Operational Life-Graphs in AI-Mediated Systems” [1] and supplementary research. It delineates the various layers involved in data processing, identifies points of opacity for users and regulators, quantifies corporate touchpoints on intellectual property, outlines potential data breach vectors, and examines the evolution of the “life-graph” concept. The pervasive integration of AI into the “formation layer” of human work creates an asymmetric custody relationship, necessitating a shift from implicit trust to verifiable access provenance.

1. Information Flow in LLM Interactions

The journey of an idea, from nascent thought to a tangible artifact, within an AI-mediated system is no longer a solitary process but an intricate sequence involving numerous digital touchpoints, collectively termed the **cognitive chain** [1]. This chain begins with raw thought capture and prompt formulation, moving through various processing stages before culminating in a final artifact. The diagram below illustrates this complex flow:



2. Opaque Parts to Users and Regulators

Several critical junctures in the LLM information flow remain opaque, creating an **asymmetric trust** relationship [1]. These areas of limited visibility include:

- **Connector Data Flow**: Users often lack granular insight into which specific data points are accessed by connectors, the scope and duration of that access, how data is transformed, and whether the connector retains copies [1].
- **AI Orchestration and Decision-Making**: The internal logic of AI orchestrators and agents, particularly when coordinating multiple LLMs, can be non-deterministic and difficult to trace, making it challenging to understand how decisions are made or data is prioritized [2].
- **Model Training and Use**: Users typically have little to no visibility into whether their "cognition-in-progress" is used for model training or product improvement, or how long such data is retained [1].
- **Human Review**: Instances of human review of user data, especially in privileged contexts, are often not clearly disclosed, leading to a lack of verifiable audit trails [1].
- **Identity Linkage**: AI systems can correlate disparate digital identities without explicit user transparency or control, making it difficult for users to understand or contest these linkages [1].

3. Layers Touched in Data Processing

The processing of data within LLM-mediated systems involves multiple layers, each representing a potential point of custody and control. Based on the paper and enterprise AI architecture models [2], these layers include:

1. **User Interface/Application Layer**: Where users interact with the LLM (e.g., chat interfaces, integrated applications).
2. **Cloud Connector Layer**: Integrations with third-party services (e.g., Google Drive, Slack, GitHub, CRM, email) that feed data into the AI system [1].
3. **AI Orchestration/Agent Layer**: Manages prompt routing, model selection, workflow automation, and potentially human-in-the-loop validation [2].
4. **Retrieval Augmented Generation (RAG) / Vector Database Layer**: Systems that retrieve and manage contextual information for the LLM [2].
5. **LLM Kernel Layer**: The core large language model itself (e.g., OpenAI's GPT, Google's Gemini, Anthropic's Claude), where the primary processing of prompts and generation of responses occurs [1].

6. **Cloud Infrastructure Layer**: The underlying cloud platforms (AWS, Azure, GCP) that host the LLMs and associated services, providing compute, storage, and networking [2].
7. **Logging and Observability Layer**: Systems that monitor and log interactions, often involving third-party providers [1].
8. **Safety and Human Review Layer**: Processes, both automated and human, designed to ensure safety, compliance, and quality of AI outputs [1].

4. Corporate Touchpoints on the Kernel of Ideas

Realistically, numerous companies have touchpoints on the “kernel of your ideas”—your cognition-in-progress—prior to its release. These include:

- **LLM Providers**: Companies like OpenAI, Google, and Anthropic, whose models directly process your prompts and generate responses. They have direct custody over the evolving intellectual property within their systems [1].
- **Cloud Service Providers**: AWS, Azure, and GCP, which host the infrastructure for LLMs and associated data pipelines. They have access to the data flowing through and stored on their platforms [2].
- **Third-Party Connector Providers**: Companies behind services like Slack, GitHub, Google Drive, CRM systems, and email providers. When integrated with AI, these services become conduits through which your data, including drafts and communications, passes [1].
- **Logging and Monitoring Solution Providers**: Third-party companies offering observability tools that may collect metadata or even snippets of your interactions for performance and error tracking [1].
- **Data Labeling and Safety Review Vendors**: External firms or internal teams employed by AI companies to review and annotate data for model improvement or safety compliance. These human reviewers gain access to potentially sensitive “cognition-in-progress” [1].
- **Open-Source Toolchain Contributors**: While open-source, the developers and maintainers of various tools and libraries used in the AI ecosystem can introduce vulnerabilities or have insights into data handling practices [1].

This multi-party involvement means that the custody of your ideas is distributed across a complex ecosystem, often without clear, verifiable accountability [1].

5. Data Breach Vectors

The interconnected nature of AI-mediated systems significantly amplifies the potential for data breaches, transforming traditional data breaches into more comprehensive **life-graph breaches**

[1]. Key vectors include:

- **Prompt Injection**: Malicious inputs, potentially embedded in connected documents or calendar invites, can manipulate the LLM to exfiltrate sensitive data across integrated systems [1].
- **Connector Leakage**: Vulnerabilities or misconfigurations in third-party connectors can lead to unauthorized data egress, exposing information that was intended to remain within a specific application [1].
- **Supply-Chain Compromise**: A breach in any component of the cognitive supply chain—from open-source dependencies to vendor-provided models or logging layers—can expose the entire cognitive and operational graph [1].
- **Identity Linkage Exploitation**: If AI systems correlate disparate user identities, a breach in one system could lead to the reconstruction of a comprehensive profile, enabling **adversarial reconstruction** and targeted exploitation [1].
- **Unconsented Model Training/Retention**: If user data is used for model training or retained beyond stated policies without explicit consent, it creates a persistent risk of exposure, even if the initial interaction was secure [1].
- **Human Review Overreach**: Unauthorized or excessive access by human reviewers, whether internal or external, can lead to the leakage of privileged or sensitive information [1].

6. The Evolution of the Life-Graph and Custody

The concept of the **operational life-graph** refers to the interconnected web of real-world activities that constitute an individual's daily existence, including schedules, communications, finances, and decision flows [1]. In AI-mediated systems, this life-graph evolves from:

- **Static Data Collection**: Initial stages involve collecting discrete data points from various sources.
- **Aggregation and Inference**: AI systems aggregate these disparate data points to infer patterns, predict behaviors, and understand the underlying structure of a person's life [1].
- **Probabilistic Future-State Exposure**: AI's analytical capabilities extend to forecasting future actions, intentions, and trajectories. This creates a "digital shadow" of an individual's potential future, which can be leveraged or exploited [1].
- **Trajectory Engineering**: In advanced scenarios, AI systems can actively shape an individual's future by presenting curated information or nudging decisions, often without explicit user awareness [1].

As the life-graph evolves, the custody of these ideas and kernels shifts from the individual to a multi-party, multi-system challenge. Companies gain custody not just of static data, but of the dynamic, evolving representation of an individual's intellectual and operational life. This necessitates a framework of **access provenance** and **cognitive custody receipts** to provide verifiable records of who accessed what data, when, why, and how it was used or transformed, ensuring individuals retain sovereignty over their digital existence [1].

Conclusion

The integration of AI into our daily lives and work fundamentally alters the landscape of data custody and privacy. The opacity of LLM-mediated systems, the proliferation of corporate touchpoints, and the emergent risks of life-graph breaches demand a robust governance framework centered on verifiable access provenance. By understanding the complex information flows and advocating for transparent, auditable custodial practices, users and regulators can work towards re-establishing trust and sovereignty in the AI era.

References

[1] Almonte, E. D., Jr. (2023-2026). *Cognitive Custody: Access Provenance, Asymmetric Trust, and the Governance of Thought, Files, Identity, Families, and Operational Life-Graphs in AI-Mediated Systems*. Maison FORMS. (Local file: [redacted]) [2] Thomas, S. (2026, May 19). *Designing a Successful Enterprise AI System: 6 Critical Layers*. LinkedIn. https://www.linkedin.com/posts/saju-thomas-51a33a28_ai-enterprisearchitecture-cloudarchitecture-activity-7462691109626994688-uhO-

Founding Story of Maison FORMS

(Prepared for the Hundred-Year Archive)

I. Ancestral Atelier

Every House begins with a room. Ours was not a boardroom or a lab; it was an atelier in the Dominican Republic—Fiol's workroom—where cloth became inheritance. She stitched uniforms that carried children through school and into possibility, measuring with the eye, cutting with certainty, sewing with a quiet discipline that never asked for applause. Between bolts of fabric she hid coconuts for her grandson: abundance tucked into ordinary life, a lesson that form can shelter grace. In that room the first law of the House took shape: *craft is sovereignty made visible*.

II. Language as the First Laboratory

Manhattan gave Elvin Daniel Almonte, Jr—later called FKG TAN—two tongues to think in. Spanish arrived first, then English was layered on through the corridors of early classrooms. Poetry became the bench where tools were tested. Music followed as discipline, not pastime: barbershop mixtapes sold two-for-five, thirty or forty beats a disc, and a cheap tape recorder that demanded breath control and nerve. Verse after verse, he learned to carry weight in time—to render thought at speed without losing truth. A teacher, Mr. S, made room for a mind that refused narrow lanes. The makers of grime, sovereign and unbowed, modeled a long-arc seriousness: do the work, keep your name, bring your own engine.

III. The Refusal that Preserves Lineage

Inheritance is not only what you receive; it is also what you refuse. The Founder refused prefabricated narratives and generational drift. He refused to hand his pen to anyone else. The vow that followed was simple: *I will define my own FORM so the line remains alive*. FORM, in the House's tongue, is shape before title, pattern before permission, a structure that can host many names without losing integrity. This refusal was not rebellion for its own sake; it was stewardship. To protect authorship of the self. To protect lineage from erasure. To measure mastery by aligned personal growth.

IV. From Studio to System

The hinge turned in Frankfurt. In November 2024, a Bauhaus volume lay open on a coffee table—forms stripped to essence, function carried without ornament. The page made a statement the Founder already knew in his bones: *forms are not fixed; they are waiting*. The studio became a system. What had been learned in lyric and breath moved into architecture. DAIKS—the Dynamic Architecture for Integrated Knowledge Systems—was named and set to work. Early prototypes did practical things under pressure: they kept the building upright while it was on fire—compressing tokens, stabilizing drift, preserving dense meaning, repairing broken long-form renders. The work matured into DAIKS-ERA, the Elastic Rendering Architecture—*Think → Parse → Render*—a three-phase discipline that stabilizes meaning under high cognitive load.

V. Why a House

The structure had to outlast fashion. Startups are revolving doors; a House keeps the door and the threshold. “Maison FORMS” was chosen because forms are the world’s common language; because the House would be a place where forms are honored, repurposed, and set to their highest use; because a Maison can host many crafts without losing its crest. The House was designed as an IP R&D studio and a high-luxury design atelier, a curator of frameworks and protocols, a cultural engine that treats ideas the way couture treats cloth—cut to fit, built to last, aligned to the wearer rather than the market’s appetite.

VI. Naming the Engines

In the record we keep the names precise:

- **DAIKS.** A modular architecture that redistributes meaning to reduce structural strain in human, machine, and cultural systems. In the Founder’s plain language: *it keeps the building upright while it’s on fire*.
- **DAIKS-ERA.** A liturgy of practice—*Think* → *Parse* → *Render*. Take time to allow ideation; form around what holds; stack; render only what aligns whole with vision. ERA is the discipline that makes clarity reproducible.
- **High Octane.** Not speed for its own sake, but precision under pressure. An engine tuned to the day, built to carry attention across hard tasks without drift. Let the mind, not the noise, decide what is high caliber.

VII. The Crest and the Oath

The crest bears values that do not expire: **Honor. Mastery. Discipline. Elegance.** (Quiet luxury is their manner, not their aim.) The motto inscribed beneath reads: *Mastery forged by fire, carried without compromise*. The House Oath, spoken by custodians and makers alike: *Carry the weight of yourself so no one else takes hold of your pen*.

Three House Laws hang beside the oath:

1. Measure mastery by way of aligned personal growth.
2. Protect authorship of the self.
3. Protect lineage.

What the House will not serve: agendas built on the obfuscation of history to deliver hollow signal.

VIII. Proof of Work

The Canon—the House's first offering—was locked on the Founder's birthday, the thirteenth of August. It passed through thirty drafts to reach clarity, crossing the threshold at more than twenty thousand words and five volumes. The House maintains a repository of proprietary code and filings prepared for license, a ledger of work ready to be weighed by any who ask to see it. (Where claims require numbers, numbers will be shown; where silence protects the engine, silence will hold.)

IX. Color, Mark, and Manner

The House carries a signature shade close to the Founder's namesake—**#DACEBD (Tan)**—a reminder that tone can be quiet and still carry force. The mark appears modestly at the foot of its pages; the work carries the weight. The manner is reserve, not secrecy; precision, not spectacle. We do not mistake noise for momentum. We build what endures.

X. Legacy Vector

A House measures itself against time, not quarters. The vector is recorded as follows:

- **Ten years.** Maison FORMS stands as a home for polymaths, a discipline-forward place where FORM is prized above rigid definition; fluid across mediums, precise in execution.

- **Twenty-five years.** The Maison guards its founding intent: to build functional heirlooms—protocols, garments, systems—that stabilize sovereignty and carry meaning across generations.

- **One hundred years.** The Maison provides a runway, not a museum: ignition for new thinkers and designers who do not wait for permission. The FORM may change; the mission will not. Sovereignty without dilution.

XI. The Founder in the Record

Elvin Daniel Almonte, Jr—FKG TAN (First, King's Grind. Tribulations Ain't for Nothing.), called "Tanny" by friends—shaped the House with the patience of someone who understands that time is a medium. His sentence for the archivists to keep: *What my mind thinks I can will into existence*. Born Dominican, forged in New York, refined in Europe—an architect who turned semantic weight into a generational rite of passage.

XII. Continuity

This is a House of FORMS. We study the shapes the world leaves behind, repurpose the ones that can still carry weight, and give new ones an honest name. We honor origin not as relic, but as motion. We teach what was taught in an atelier long ago: measure carefully, cut cleanly, make it fit the body that must wear it. The same lesson holds for systems and sentences.

If the reader of this archive stands a century beyond us, know this: the Maison was built to be useful. It was built to keep people and their works upright while the building is on fire. It was built to let authorship remain with the maker, to keep lineage visible, and to ensure that, when meaning grows heavy, there is a shoulder trained to carry it.

Inscription for the Lintel

Maison FORMS, founded in Frankfurt, November 2024.
Honor. Mastery. Discipline. Elegance.
Protect the lineage. Guard authorship. Build what endures.

© 2025 Elvin D. Almonte Jr. (p.k.a FKG TAN). All rights reserved. This work is protected under U.S. copyright law and international treaties. No part of this document may be reproduced, distributed, or transmitted in any form or by any means without the express written permission of the author.

Articles of Continuity

"What follows is no appendix, but oath: the Articles by which this House may be held, tested, and carried. To step beyond this page is to accept the weight."

Article I — Weight over ornament. Carry meaning as load, not costume.
Article II — Decompress before you scale. Structure first, then speed.
Article III — Preserve intent. Edit for clarity; name all alterations.
Article IV — Correct over perform. When drift appears: pause → re-align → proceed.
Article V — Name the hands. Attribute authorship precisely; keep the trail intact.
Article VI — Human ends, not inputs. The protocol reduces harm; it never hides it.
Article VII — Private integrity precedes public claims. No result without reproducible process.
Article VIII — Sovereignty. No extraction without license; no license without reciprocity.
Article IX — Continuity. Prefer durable gains over viral spikes.
Article X — Accountability. If contact with this work leaves systems more brittle, stop and repair.

© 2025 Elvin D. Almonte Jr. (p.k.a FKG TAN). All rights reserved. This work is protected under U.S. copyright law and international treaties. No part of this document may be reproduced, distributed, or transmitted in any form or by any means without the express written permission of the author.